2015

Food Inspection Violation, Anticipating Risk (FIVAR) Montgomery County, MD

A REPORT BY OPEN DATA NATION

CAREY ANNE NADEAU, FOUNDER & CEO

& SOFIA HEISLER, DATA SCIENCE CONSULTANT

SUMMARY

From November 12, 2013 to November 12, 2015, the Montgomery County Department of Health and Human Services inspected 3,968 food establishments in Montgomery County, Maryland in order to uncover issues that might affect the health and safety of the patrons who visit these establishments. [1] Of all inspections over this three-year period, approximately 30% revealed violations, such as contaminated food and improper refrigeration that could have caused the outbreak and spread of foodborne illness.

In Montgomery County, inspectors are deployed to food establishments with a list prioritized according to state and national inspection standards.[2] Within 24 hours of the inspection, the results are uploaded to the County's online open data portal (www.data.montgomerycountymd.gov) and made publically available. The inspection results then remain publically available in the open data portal for three years thereafter.

The 'Food Inspection Violation, Anticipating Risk' (FIVAR) model developed by Open Data Nation, puts this open data to good use. Using historical data of inspection results, food establishment characteristics, neighborhood characteristics, weather and time of the year, the FIVAR model prioritizes Montgomery County food establishments by their risk of a violation.

Based on a two-month evaluation, the data-driven FIVAR model anticipates that Montgomery County can identify 49 additional violations (27% more) in the first month; on average, violations can be identified at least 3 days earlier than business as usual.[3] This

[1] The Licensure & Regulatory Services Program inspects all licensed retail food establishments in Montgomery County for a variety of reasons (e.g. obtaining a permit, regular check-ups, or in response to complaints.) Included in this overall surveillance are two types of inspections that are conducted on a routine basis. The first type, a comprehensive inspection, is a thorough inspection that evaluates the sanitation, maintenance, and food service operations for the facility. This inspection includes many of the items found in the monitoring inspection, such as critical temperatures and food handling procedures. The second type, a monitoring inspection, involves checking the food service operation for critical food temperatures, equipment temperatures, and general food handling / cleanliness practices. While monitoring inspections are not as detailed as environmental inspections, they help ensure that the facility is operating safely throughout the year. (https://data.montgomerycountymd.gov/Health-and-Human-Services/Food-Inspection/5pue-gfbe)

[2] The frequency of inspections is based on the food-borne illness risks associated with the food being processed at a facility. Based on the risk each facility is designated as high, medium, or low priority. High priority risk facilities prepare food products a day or more in advance or utilize any combination of two or more processes such as a cooking, cooling, reheating and hot holding food over four hours. Moderate priority risk facilities prepare and cook food products which are served immediately or within four hours. A low priority risk facility serve prepackage foods that are not potentially hazardous.

For high-risk facilities, a monitoring inspection is conducted at least twice a year, and moderate risk facilities are inspected at least once a year. High and moderate risk facilities also receive at least one comprehensive inspection every year. Low risk priority facilities typically receive one comprehensive inspection every two years. (C) Denotes a Critical Violation. A Critical Violation means a food safety requirement that requires immediate correction. Failure for immediate correction results in cessation of some or all food operations or closure of the facility until violation is able to be corrected. (https://data.montgomerycountymd.gov/Health-and-Human-Services/Food-Inspection/5pue-gfbe)

[3] The number of additional violations (49) is specific to a two-month trial based on 403 randomized inspections. This number will increase when the trial is extended over a longer period and includes more violations. For example, if we assume that the model will identify 10% of all establishments eligible for inspections as high-risk for a violation, when

means that with FIVAR, Montgomery County can deploy inspectors more efficiently and effectively, reduce exposure to unsanitary conditions, and prevent the spread of foodborne illnesses.

While the model in isolation is informative, to accrue benefits to Montgomery County, it must be incorporated into the operations of the Montgomery County Department of Health and Human Services. To this end, Open Data Nation proposes a pilot project to evaluate the FIVAR model that aims to understand the existing systems and processes of food inspections and opportunities to incorporate the FIVAR model, such as providing real-time prioritized inspection task lists to inspectors.

BACKGROUND

At least 450 municipalities in the United States publish the results of food inspection violations online. In most municipalities, this data contains information about the establishment's name, address, inspection date, and results.

While data is broadly available, it is only occasionally analyzed to inform or evaluate the operations of health inspectors. In 2014, in an open-source project, the Chicago Department of Public Health developed a data-driven model that aimed to use food inspection results to predict more violations and sooner (http://chicago.github.io/food-inspections-evaluation/). From the open-source results of the model developed for this project, we know that it is theoretically possible to use data to identify 25% more critical food inspection violations and more than a week earlier, on average.

With the Chicago example in hand, Montgomery County, Maryland, set out to adapt the model to local data availability and determine whether similar benefits could be achieved in their municipality. Working with the Office of the Chief Innovation Officer of Montgomery County, Open Data Nation, a social benefit company that provides data science consultation and open data training, cleaned and compiled data from Montgomery County's open data portal and other sources. With this data, Open Data Nation fit the FIVAR model that measures the theoretical benefits possible in Montgomery County.

The FIVAR model and the model developed in Chicago differ in important ways. For one, data are not consistently available or reported consistently between the two municipalities. For example, the age of each establishment was available and incorporated into the model in Chicago, but it was not readily available to use for the FIVAR model in Montgomery County. In addition, the values reported are for different periods of time. In Chicago, the data described inspections over a period of five years (2011 to 2015) whereas three years of data (2013 to 2015) were available in Montgomery.

4,000 observations are included (the approximate number of inspections in one year), we would anticipate identifying 400 high-risk establishments in the first month, with diminishing returns over time.

Secondly, local conditions and feedback from inspectors led to the inclusion of different variables. For example, in Montgomery County, officials suggested that construction might be an important determinant of violations, as this disturbs the ground and pests. The inclusion of nearby construction permits in the FIVAR model for Montgomery County, unlike Chicago, ultimately was predictive and improved the model.

The remainder of this report provides a more in-depth description of the FIVAR model, including a description of the data, method, evaluation, limitations, and next steps.

DATA

Montgomery County health inspection data from October 2013 through November 2015 were used to develop and test the model. In addition, the following types of data pertaining to Montgomery County were used as inputs to the modeling:

- Address-level data pertaining to complaints to 311
- Crime records
- Demolition and construction permit data
- Liquor licenses
- Weather data
- Yelp data (number of reviews, star rating, and cuisine)

In developing the FIVAR model, Open Data Nation sought to predict the likelihood that an inspection would result in any type of violation. The following variables were found to be predictive of the probability of a violation being identified and incorporated into the final version of the model (for a complete list, see the Appendix 1):

Food establishment's characteristics:

- Past failures: Previous violations are strongly associated with an increased risk of new violations;
- Possession of a liquor license: Food establishments with a liquor license are statistically more likely to have violations identified;
- Category of food establishment: Restaurants and take-out food joints were found to be significantly more likely to have violations than all other types of establishments, while schools were found to be significantly less likely to have violations;
- Location of the food establishment: food establishments located in Silver Spring were found to be more likely to incur violations, while Rockville, Bethesda, and Olney were associated with a lower risk of violations;
- First inspection: Violations were significantly less likely to be found on the initial inspection than subsequent inspections;
- Time since previous inspection: The probability of a violation being found increased with time since the preceding inspection;
- Number of Yelp reviews: The number of reviews an establishment has received on Yelp was positively associated with the probability of a violation;

- Cuisine type as reported on Yelp: Some of the more common types of cuisines, such as American, Asian, Middle Eastern, and Mexican places, as well as bars, were found to have a higher probability of violations than others. Dessert places (donut, cupcake, and chocolate shops) were found to have a lower probability of violations.

Neighborhood characteristics

- Burglary rates: The probability of violations being identified is positively related to burglaries in the area within the 90 days leading up to the inspection;
- Larceny rates: the probability of violations being identified is negatively related to larcenies in the area within the 90 days reading up to the inspection;[4]
- Illegal dumping complaints: The probability of violations being identified is weakly but positively related to complaints to 311 classified as "Illegal dumping" recorded in the 30 days leading up to the inspection;
- Other health code violations in the area: the probability of violations being identified is positively related to other health code violations recorded in the area within the preceding 90 days;
- 311 complaints: the probability of violations being identified is weakly but positively related to complaints to 311 in the 90 days leading up to the inspection.

Weather and time of year

- Temperature: Violations were more likely to be identified on days with hotter maximum temperatures on the day of the inspection;
- Humidity: Violations were more likely to be identified on days with higher humidity on the day of the inspection;
- Month in which inspection was performed: Inspections performed in August were significantly more likely to identify violations, while inspections performed in May were significantly less likely to do so.

[4] Presumably, crime does not by itself *create* conditions for health code violations, but prevalence of certain types of crime may be indicative of some characteristics of an area. For example, one possible explanation for the relationship between larceny, burglary, and health code violations is that larceny, defined as non-violent theft such as pickpocketing, may be more prevalent in busy, high-traffic areas such as malls or movie theaters, while burglaries are more likely to occur in low-traffic areas or lower-income areas. The characteristics of food establishments associated with each type of area are likewise likely to be different.

METHOD

Two separate logistic models were fitted to data between October 12, 2014 and August 31, 2015, using stepwise variable selection. Depending on the availability of Yelp data for a food establishment, one of the two models is applied to evaluate the probability of identifying a violation on a particular day. The coefficients associated with the two models are available in Appendix 2.

The first model was fitted on data associated with institutions for which valid Yelp data (e.g., cuisine, number of stars, and number of reviews) could be obtained. These tend to be popular restaurants, as well as a small subset of other business types (e.g., country clubs).

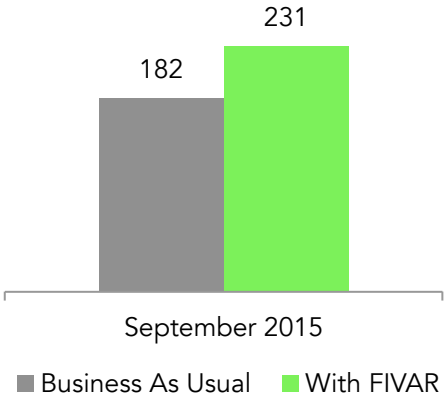
The second model was fitted on data for institutions that had missing or incomplete Yelp data. These tend to be non-restaurants (e.g., schools or gas stations), as well as restaurants that get minimal foot traffic.

MODEL EVALUATION

In order to test the effectiveness of the model, the number of violations anticipated by the model for September and October 2015 was compared to the results of inspections, according to business as usual. The goal of the test was to evaluate how many violations would have been identified sooner, and how much sooner, had inspections been prioritized according to the FIVAR model.[5]

Over the two-month period from September to October 2015, 403 food inspection violations were reported.[6] Of these, 182 violations (45%) were identified in September and the remaining 221 (55%) were identified in October.

Had the same inspections been prioritized according to the FIVAR model, 231 (57%) of the 403 violations could have been identified in September. In other words, 49 more violations (27% more) would have been found sooner, in the first month of the trial. On average, the FIVAR model identified the same violations 3 days sooner than business as usual.



Looking at the prioritized list of establishments resulting from the FIVAR model, the 50 establishments ranked most likely to have a violation were nearly two times as likely to have reported violations (60% of the time) over the two-month period than the average rate of all violations (33%) over the same period.

[5] The test is similar in scope and scale to that which was conducted in Chicago. (http://chicago.github.io/food-inspections-evaluation/)

[6] Count in data as of November 9, 2015

LIMITATIONS

Previous research on predicting health code violations in Chicago suggests that other variables may be predictive of health violations, but were not available for this analysis. The following variables may have further improved the model:

- **Inspector identifier.** Montgomery County records inspector coverage areas, but this does not allow an individual inspection to be mapped to the inspector responsible. Variability between inspectors was highly predictive in Chicago and would most likely improve the FIVAR model in Montgomery County.
- **Length of time establishment has been operating.** Montgomery County records length of establishment in business licensing data, but this data were not available in a machine-readable form that would have enabled its inclusion in the FIVAR model at the time of this report.

NEXT STEPS

To test whether the theoretical benefits of the FIVAR model may be achieved in Montgomery County, the model would need to be incorporated into practice and its impact evaluated.

The FIVAR model could be used to prioritize work order task lists for inspectors, as frequently as the data required for the model is published. To this end, data that were manually constructed or requested for this one purpose would need to be automated and ideally available from a set point, rather than on a rolling basis. These data include:

- **Complete list of all establishments eligible for health inspections.** A list of all eligible establishments will ensure that as new establishments come on line, they are included in the model.
- **Illegal dumping 311 complaints.** These data were obtained separately from Montgomery County officials, per request, and would need to be made consistently available.
- **Liquor license machine-readable data.** The liquor license data are currently not available publicly in a machine-readable format. Moreover, there is no business identifier that would permit automated matching between liquor license data and the food inspection data, which means that the matching process has to be performed manually with any changes in liquor licenses. In order to ensure that up-to-date and accurate data are available to feed into the model, the liquor license data should be made available in machine-readable form and include a business identifier that would allow them to be matched to health inspections data

An experiment to test the FIVAR model could be modeled after the pilot project used to test the food inspection algorithm in Chicago. The impact of the FIVAR model would be measured as the difference in number and speed to identification between an experimental group of inspectors operating with the FIVAR model and another group of inspectors operating without the model.

CONCLUSION

The Food Inspection Violation Anticipating Risk (FIVAR) model developed by Open Data Nation on behalf of Montgomery County, Maryland puts open, public data to good use. The results of this algorithm suggest that Montgomery County can more strategically deploy their limited resources of food inspectors and improve the public health and safety of patrons at food establishments. As one of only a few locations to have created a predictive model that prioritizes food establishments according to their likelihood of a violation, Montgomery County is uniquely positioned to again lead the open data movement by making data-driven decisions to save money and deliver high-impact public benefit.